Scim: Intelligent Skimming Support for Scientific Papers

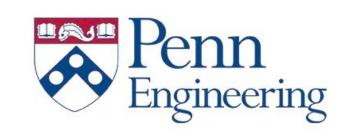
Raymond Fok (presenter), Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti A. Hearst, Andrew Head, Daniel S. Weld





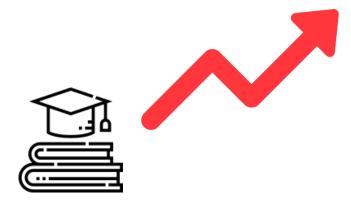
IUI 2023







The Explosion of Scientific Literature



Rate of academic publishing increased ~4% per year, over the past decade.





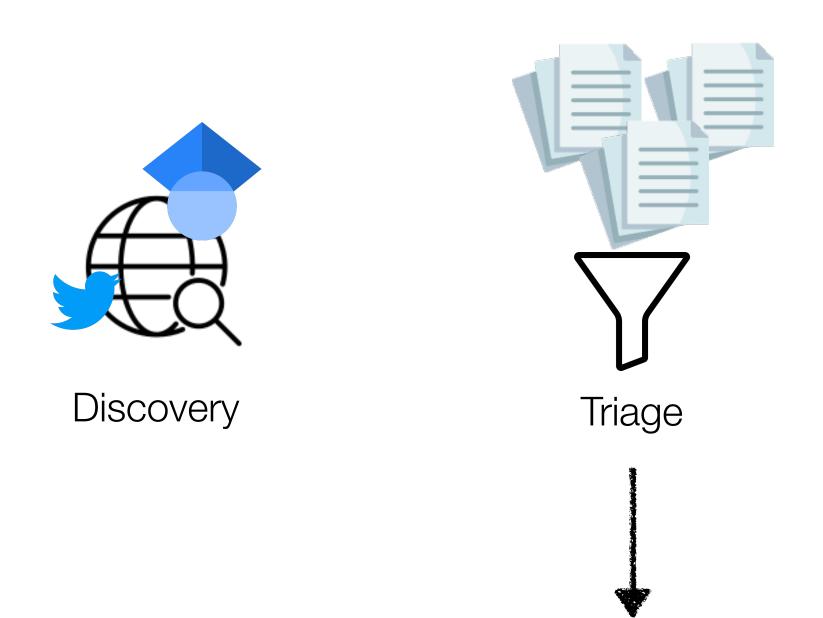


Semantic Scholar

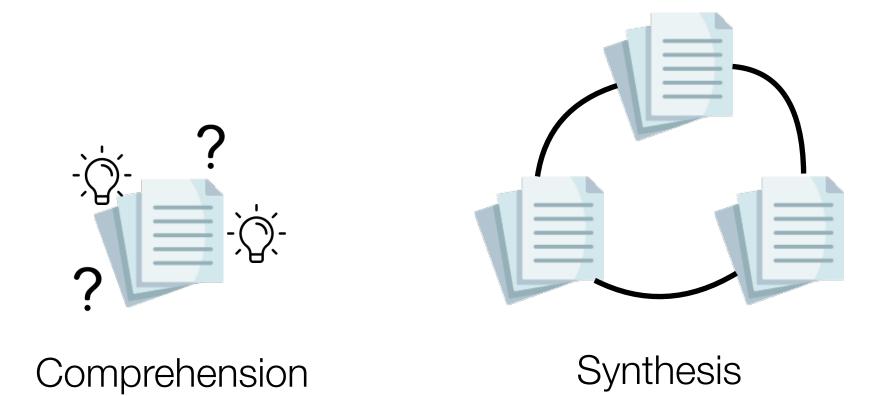
15k+ submissions per month

200M+ papers across all of science

Consuming Scientific Literature



How can we support scholars in triaging the literature more efficiently?





Augmenting Static PDFs

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova Google AI Language

{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output laver to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial taskspecific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., as question answering, where it is crucial to incor-2018a; Radford et al., 2018; Howard and Ruder, porate context from both directions. 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the re-BERT alleviates the previously mentioned unidilationships between sentences by analyzing them rectionality constraint by using a "masked lanholistically, as well as token-level tasks such as guage model" (MLM) pre-training objective, innamed entity recognition and question answering, where models are required to produce fine-grained masked language model randomly masks some of output at the token level (Tjong Kim Sang and the tokens from the input, and the objective is to De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning all pretrained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-toright architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying finetuning based approaches to token-level tasks such

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers. spired by the Cloze task (Taylor, 1953). The predict the original vocabulary id of the masked

Augmentation through automatic highlighting

20 May 2 CL] [cs 0 .04805 810

N

6

CS 2 810.04805

X

ar,

2019

May

24

CL]

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova Google AI Language

{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial taskspecific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

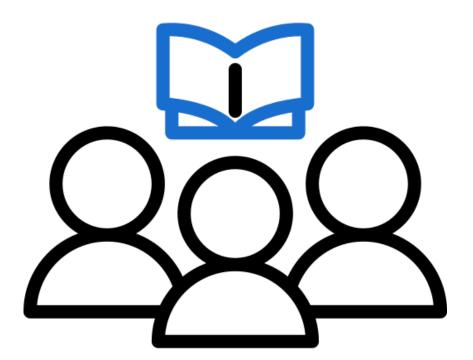
Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as guage model" (MLM) pre-training objective, innamed entity recognition and question answering, where models are required to produce fine-grained masked language model randomly masks some of output at the token level (Tjong Kim Sang and the tokens from the input, and the objective is to De Meulder, 2003; Raipurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning all pretrained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-toright architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying finetuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers, BERT alleviates the previously mentioned unidirectionality constraint by using a "masked lanspired by the Cloze task (Taylor, 1953). The predict the original vocabulary id of the masked

What do scholars want from a skimming aid?

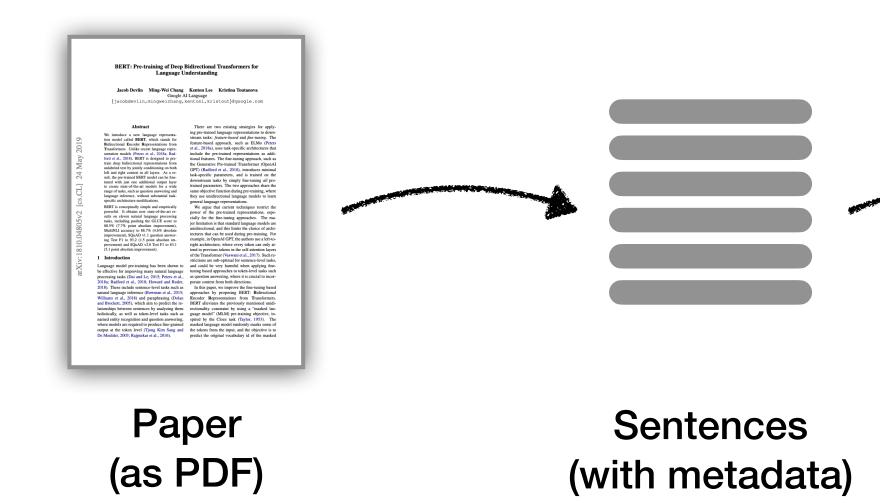


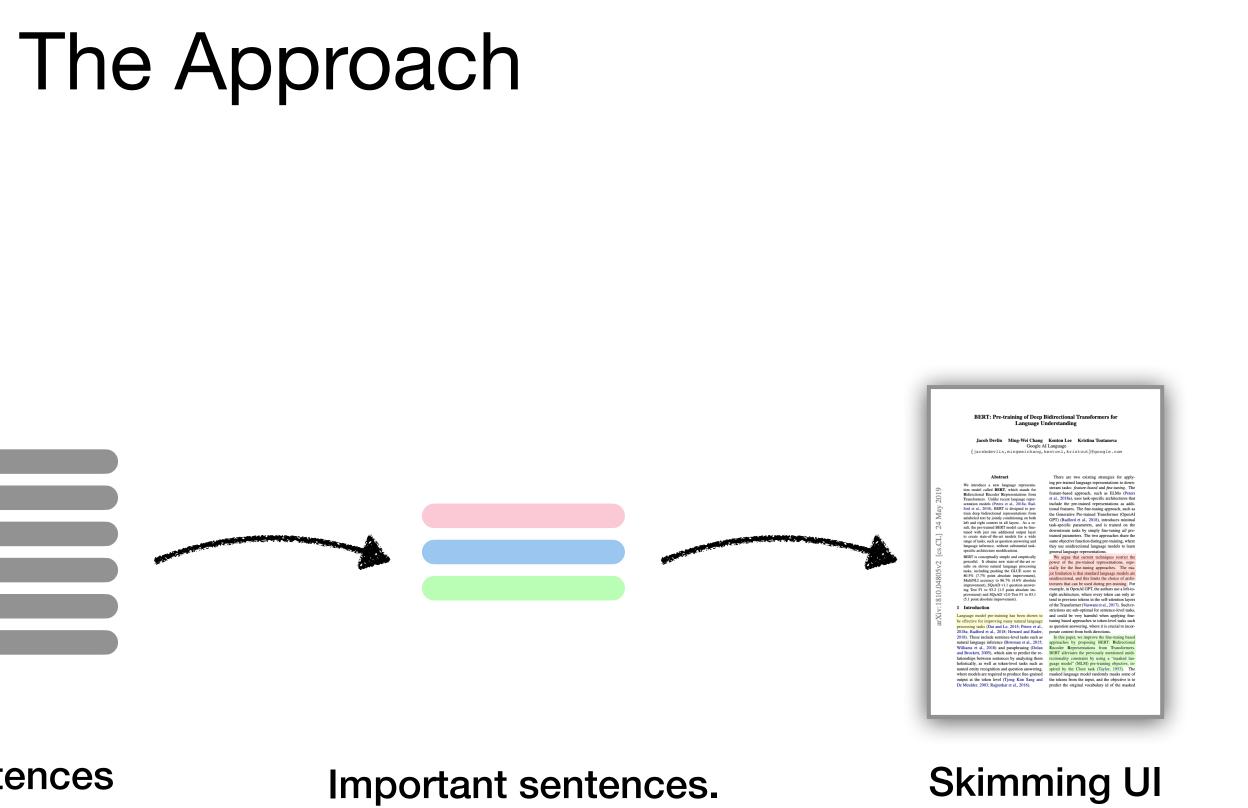
Formative user study with 10 scholars

Design Motivations

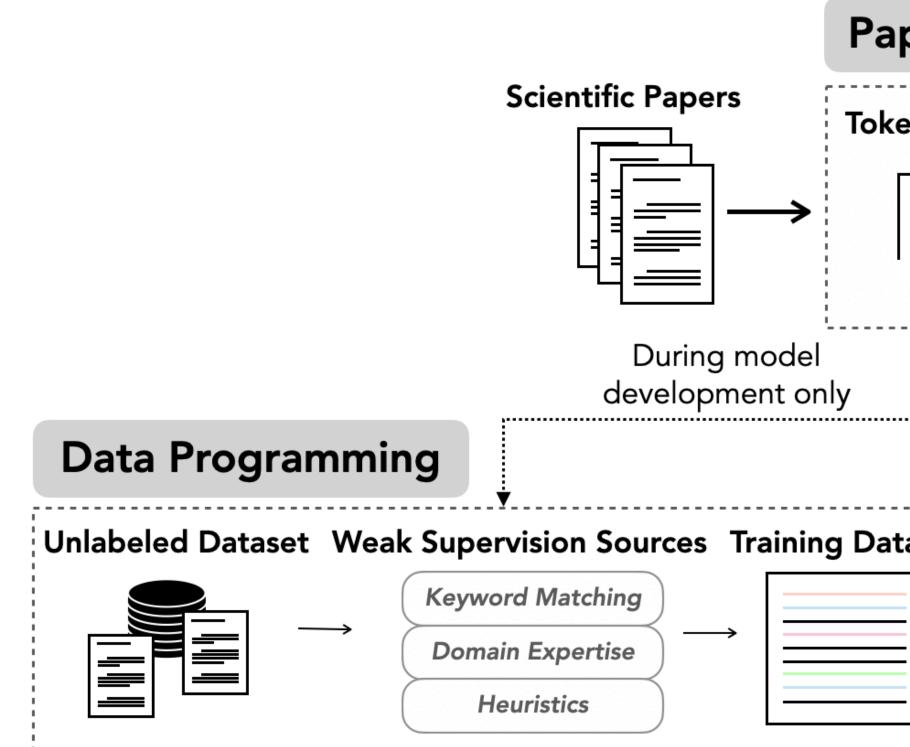
An intelligent highlighting-based skimming aid should...

- Augment readers' skimming practices.
- Highlight diverse kinds of content corresponding to information needs.
- Support skimming through text-dense sections.
- Provide user customization





The Paper Processing Pipeline



per Co	nten	t Extraction		
en Localiz	ation	Sentence Extraction	Metadata Taggin	g
_	٦		Section	
	-		Paragraph	
	•		Entity Type	
	[
		Sentence Class	sification	
aset	Facet Class		sification	Sentence Prioritizatio
		Objective	Novelty	0.9
		Result	Method	0.7
	į			

on 95 70

Aligning Highlights to Information Needs

What are their key results?

OBJECTIVE

Given a cluster of topically related news documents, we aim to detect misinformation at both document level and a more fine-grained event level.

RESULT

Experimental results show that our proposed method significantly outperforms existing methods by up to 7 F1 points on this new task.

Readers often share some common information needs when skimming.

What's new about this work?

Scim identifies **important** sentences, and then labels them with common information facets: Objective, Novelty, Result, and Method.

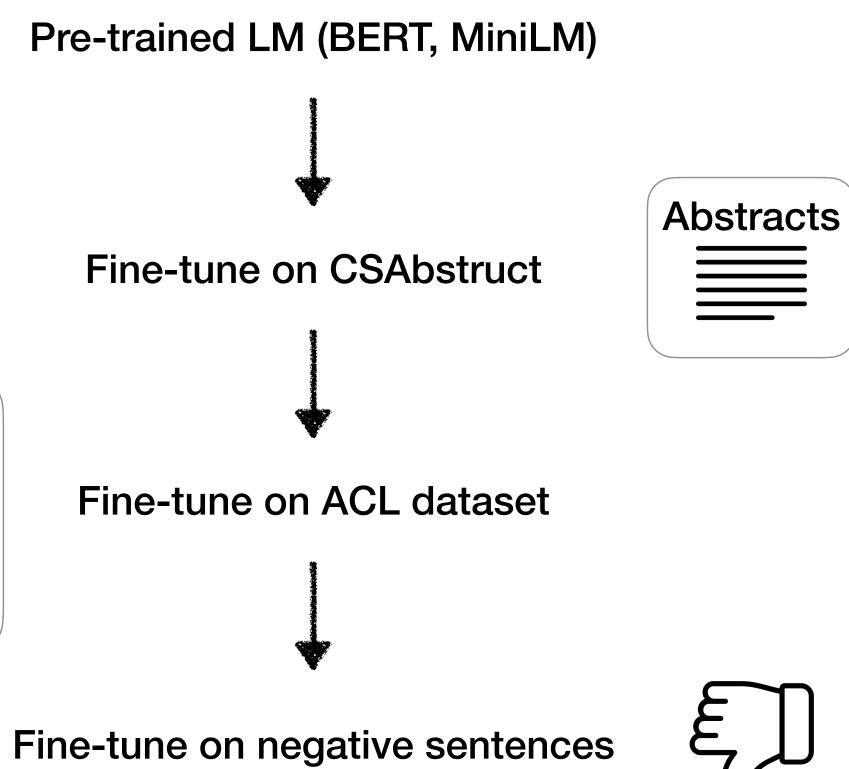
NOVELTY

However, to the best of our knowledge, no published work has considered using cross-document inference for misinformation detection.

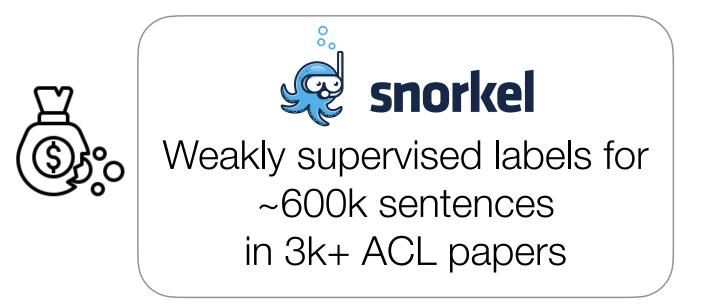
METHOD

We train a generator that generates a document from a knowledge graph (KG), and feed manipulated KGs into the generator to generate fake news documents.

Classifying Faceted Information



Data Programming



107 (1 of 9)	Objective Novelty Method Result	
	tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre- trained language models to further boost their performance. Existing works of knowledge in- fusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which di- rectly incorporates knowledge-generated atten- tion maps into the self-attention mechanism. It requires only a few extra parameters and sup- ports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent im- provements on various academic datasets for natural language understanding. It also out- performs other state-of-the-art methods which conduct knowledge infusion into transformer- based architectures. Moreover, we apply our	good chance to scores as well as applications. Recently, there incorporating kri- tectures. ERNIE PLER (Wang et textual corpora an resentation mode work. They need injecting new kn can not benefit points. K-Adapt additional neural of knowledge. It
	model to an industry-scale ad relevance applica- tion and show its advantages in the real-world scenario.	In this paper, self-attention m knowledge to ad
pu ra ac te N et	Introduction anguage models pre-trained by a large text cor- us have shown superior performances on a wide ange of natural language processing tasks. Many dvanced models based on the transformer archi- ectures achieve state-of-the-art results on various LP benchmarks. Existing literature (Jawahar t al., 2019; Hewitt and Manning, 2019) shows hat pre-training enables a model to capture syntac-	First, we show a which motivates uself-attention means in the self-attention means is to judge if the orelevant. As shows if is this pair and does not underst which rarely appropriate, "glipizide"

tic and semantic information in the self-attention

mechanism. However, the attention maps, which

o improve the quality of attention as the performance of downstream

ere have been multiple attempts for knowledge into transformer archi-NIE (Zhang et al., 2019) and KEet al., 2019) utilize both large-scale and knowledge graphs to train a repodel in a multi-task learning frameed to be retrained from scratch when knowledge, which is inefficient and it from existing pre-trained checkpter (Wang et al., 2020) integrates ral models to capture different kinds It enables adaptation based on pre-

portant information

	Disable ski	imming
	Objective (2)	
	Novelty (13)	
	Method (38)	
	Result (19)	
	well as the performance of downstream applications.	
×	In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above.	
	First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism.	
	To address the above motivation, we propose a novel architecture, namely KAM- BERT (Knowledge-assisted Attention Maps for BERT).	
	Specifically, we consider three kinds of semantic knowledge to guide the self- attention mechanism, i.e., entity, phrase	

107 ((1 of 9) Objective Novelty Method Result	
	tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre- trained language models to further boost their performance. Existing works of knowledge in- fusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which di- rectly incorporates knowledge-generated atten- tion maps into the self-attention mechanism. It requires only a few extra parameters and sup- ports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent im- provements on various academic datasets for natural language understanding. It also out- performs other state-of-the-art methods which conduct knowledge infusion into transformer- based architectures. Moreover, we apply our model to an industry-scale ad relevance applica- tion and show its advantages in the real-world scenario. 1 Introduction Language models pre-trained by a large text cor- pus have shown superior performances on a wide range of natural language processing tasks. Many advanced models based on the transformer archi- tectures achieve state-of-the-art results on various NLP benchmarks. Existing literature (Jawahar et al., 2019; Hewitt and Manning, 2019) shows that pre-training enables a model to capture syntac-	good chance to in scores as well as the applications. Recently, there he incorporating know tectures. ERNIE PLER (Wang et al. textual corpora and resentation model work. They need to injecting new know can not benefit from points. K-Adapter additional neural me of knowledge. It en trained language re instruct the self-at introduces a relative to the original model In this paper, we self-attention meet knowledge to addre First, we show a fait which motivates us self-attention meet knowledge if the que relevant. As show sifies this pair as if does not understant which rarely appear fact, "glipizide" is
	tic and semantic information in the self-attention	related to the word

mechanism. However, the attention maps, which

to improve the quality of attention as the performance of downstream

here have been multiple attempts for knowledge into transformer archi-NIE (Zhang et al., 2019) and KEet al., 2019) utilize both large-scale a and knowledge graphs to train a repodel in a multi-task learning frameed to be retrained from scratch when knowledge, which is inefficient and fit from existing pre-trained checkapter (Wang et al., 2020) integrates ral models to capture different kinds b. It enables adaptation based on pre-

f a Summarize highlights in a sidebar view

=+

		able skimming
	Objective (2)	
	Novelty (13)	
	Method (38)	
	Result (19)	
	well as the performance of downstream applications.	
×	In this paper, we propose a novel and generic self-attention mechanism enhance by explicit knowledge to address problem mentioned above.	
	First, we show a failure case of query-ad matching, which motivates us to inject	٦
	explicit knowledge into self-attention mechanism.	
	To address the above motivation, we propose a novel architecture, namely KA BERT (Knowledge-assisted Attention Ma for BERT).	
	Specifically, we consider three kinds of semantic knowledge to guide the self-attention mechanism, i.e., entity, phrase	

pu ra ac te N	anguage models pre-trained by a large text cor- us have shown superior performances on a wide ange of natural language processing tasks. Many dvanced models based on the transformer archi- ectures achieve state-of-the-art results on various ILP benchmarks. Existing literature (Jawahar	good chance to scores as well a applications. Recently, the incorporating k tectures. ERN PLER (Wang e textual corpora resentation mo work. They nee injecting new k can not benefit points. K-Adap additional neura of knowledge. trained languag instruct the self introduces a rel to the original r In this paper self-attention r knowledge to a First, we show a which motivates self-attention m map of a query is to judge if the relevant. As sh sifies this pair does not under
th	t al., 2019; Hewitt and Manning, 2019) shows nat pre-training enables a model to capture syntac- c and semantic information in the self-attention	which rarely ap fact, <i>"glipizide</i> related to the w

mechanism. However, the attention maps, which

to improve the quality of attention as the performance of downstream

here have been multiple attempts for knowledge into transformer archi-NIE (Zhang et al., 2019) and KEet al., 2019) utilize both large-scale and knowledge graphs to train a repodel in a multi-task learning frameeed to be retrained from scratch when knowledge, which is inefficient and fit from existing pre-trained checkapter (Wang et al., 2020) integrates ral models to capture different kinds It enables adaptation based on preage models. However, it does not elf-attention mechanism directly and elatively large number of parameters l model.

(=+

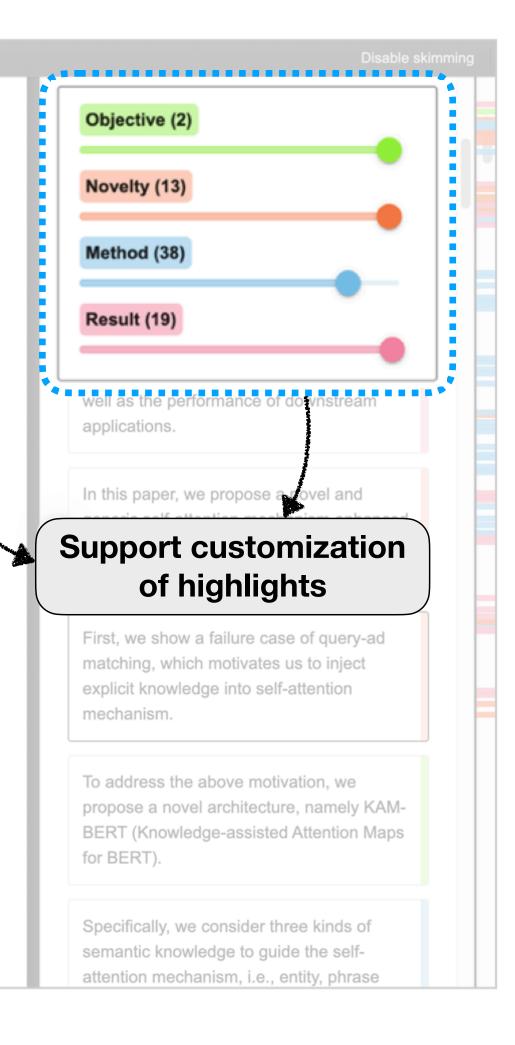
	Objective (2)
	Novelty (13)
	Method (38)
	Result (19)
	well as the performance of downstream applications.
×	In this paper, we propose a novel and generic self-attention mechanism ennanced by explicit knowledge to address problems mentioned above.
	Show distribution of highlights at a glance
	explicit knowledge into self-attention mechanism.
	To address the above motivation, we propose a novel architecture, namely KAM- BERT (Knowledge-assisted Attention Maps for BERT).
	Specifically, we consider three kinds of

107 (1 of 9		Objective				
	me are exp to tra per fus fra lar, cor and rec tio rec por is a pro nat per cor bas mo tio sce 1 In Langu pus ha range c	tion scores be echanism, are e learned impli- plicit semanti- infuse explici- ined language rformance. Ex- sion largely de meworks, whi ge-scale re-tra- nsidered. In t d generic soli- ctly incorporation n maps into the quires only a rts efficient fin- added. KAM- ovements on a tural language rforms other se- nduct knowle sed architectu- odel to an indu- n and show it enario. troduction age models twe shown su- of natural language to an indu- tage models twe shown su- of natural language to an indu- and show it enario.	sometimes icitly witho ic knowledg t external kn e models to kisting work epend on m ich are ineff aning when his paper, w ution, KAM tes knowledg te self-attent few extra pa ne-tuning on BERT achie various acad e understan state-of-the-a dge infusion stry-scale ad s advantage pre-trained perior perf nguage pro based on th ate-of-the-a function perf nguage pro based on th ate-of-the-a Existing	ineffective ut the guida ge. Thus, we howledge in further boo s of knowled hulti-task led ficient and re new knowl re propose a l-BERT, wh ge-generated ion mechan rameters are ce new knowled eves consist demic datast ding. It all art methods h into transf ver, we appled relevance a s in the real formances cessing tast he transfor art results of literature nning, 201	as they ance of we aim no pre- st their edge in- earning require edge is a novel nich di- d atten- nism. It nd sup- wledge tent im- sets for so out- swhich former- ply our applica- l-world ge text cor- on a wide sks. Many mer archi- on various a (Jawahar 19) shows	good chance to scores as well a applications. Recently, the incorporating k tectures. ERN PLER (Wang et textual corpora a resentation mod work. They need injecting new k can not benefit points. K-Adap additional neura of knowledge. I trained languag instruct the self introduces a rela- to the original m In this paper self-attention m knowledge to a First, we show a which motivates self-attention m map of a query- is to judge if the relevant. As sh sifies this pair does not under
	-	l semantic in		-	-	fact, <i>"glipizide</i> related to the w

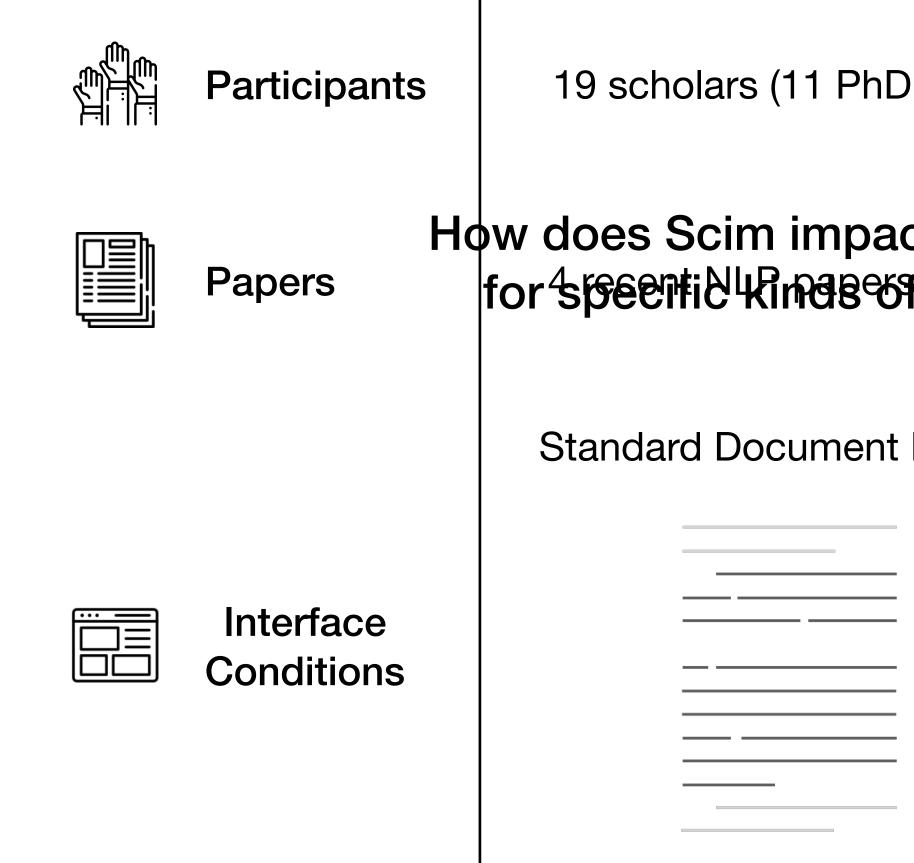
mechanism. However, the attention maps, which

to improve the quality of attention as the performance of downstream

here have been multiple attempts for knowledge into transformer archi-NIE (Zhang et al., 2019) and KEet al., 2019) utilize both large-scale and knowledge graphs to train a repodel in a multi-task learning frameeed to be retrained from scratch when knowledge, which is inefficient and fit from existing pre-trained checkapter (Wang et al., 2020) integrates ral models to capture different kinds 🚪 It enables adaptation based on preage models. However, it does not lf-attention mechanism directly and elatively large number of parameters 📩 🖬 🖡 l model.



Study 1: Usability Study



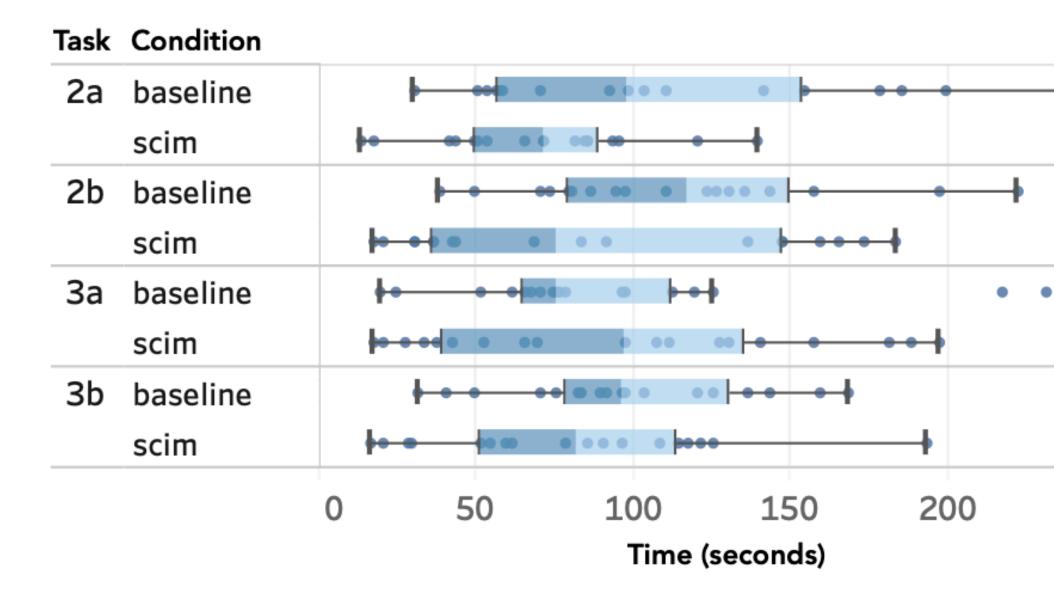
19 scholars (11 PhD, 5 Master's, 2 SWE, 1 industry researcher)

How does Scim impact readers' ability to search for specific Kinese information in a paper?

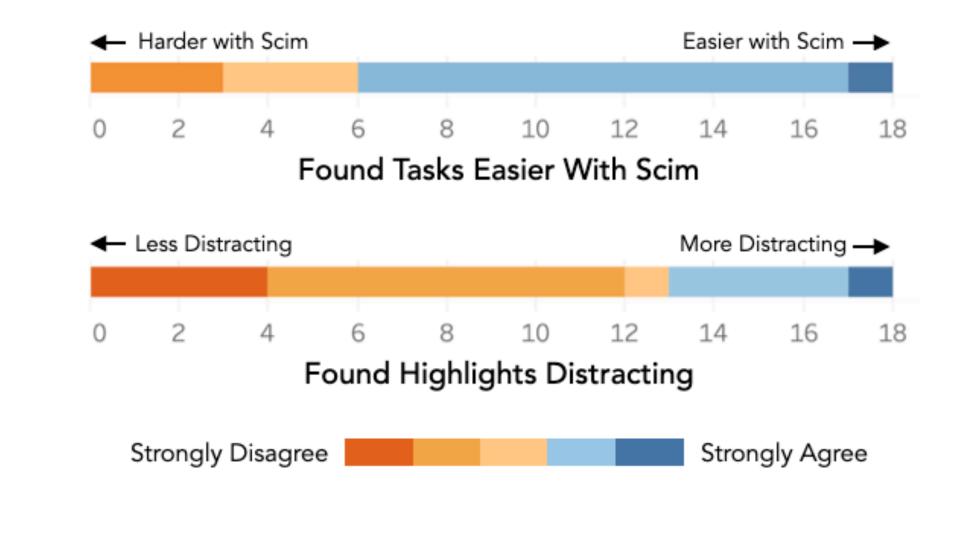
t Reader	Scim			
	(answer in highlights)	(answer not in highlights)		
_				
_				
-				
_				

Usability Study - Key Findings

Overall, participants completed informationseeking tasks more quickly with Scim than with a standard document reader.

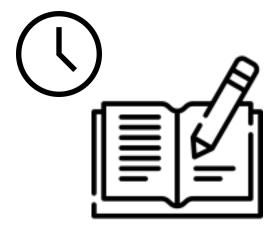


Participants felt information-seeing tasks were easier to complete with Scim, and highlights were not overly distracting.





Study 2: Diary Study



Participants skimmed at least one paper per day, and created a diary entry.

Diary entry prompts:

- - you better skim this paper.

We conducted a **2-week long diary study**

with **12 scholars** (11 PhD, 1 Master's students)

skimming papers of their choice from the proceedings of NAACL '22 How would scholars use Scim for more realistic skimming tasks?

1. Did highlights help you skim this paper? Explain. 2. List one or more ways the system could have helped



Scim's highlights provided an approachable, low-effort summary of high-value information in a paper. Readers could skim only the highlights, yet easily gain more context on-demand.

- Skimming dense texts
- Skimming papers from unfamiliar domains
 - Skimming with low engagement

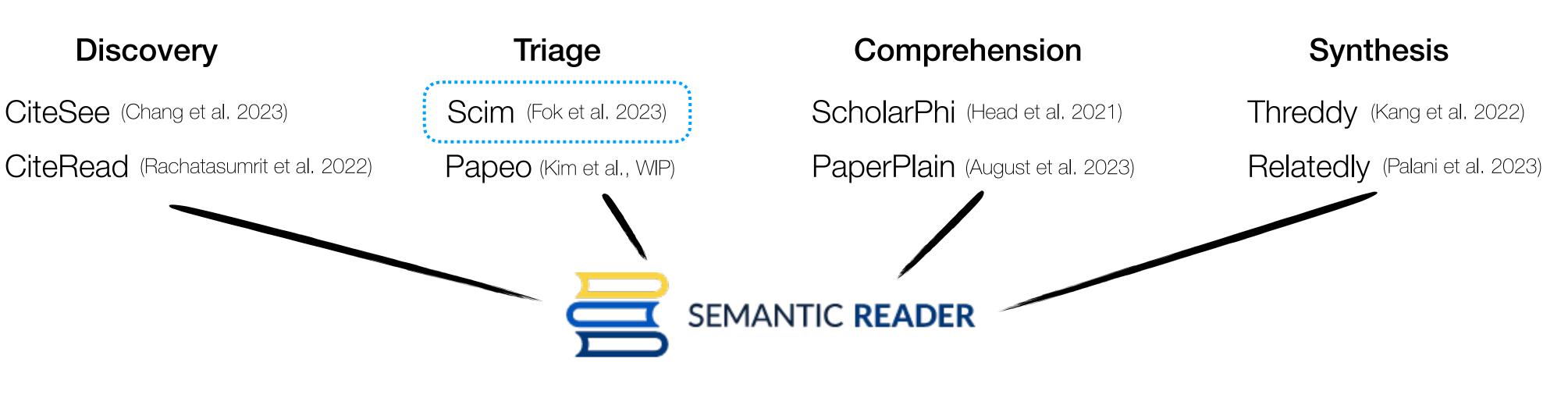


Over the two weeks, participants noted learning how to better skim with an imperfect Al assistant, i.e., with highlights that were potentially inaccurate or irrelevant.

Diary Study - Key Findings

Participants noted Scim was particularly useful for:

Al-Augmented Scientific Reading and Writing



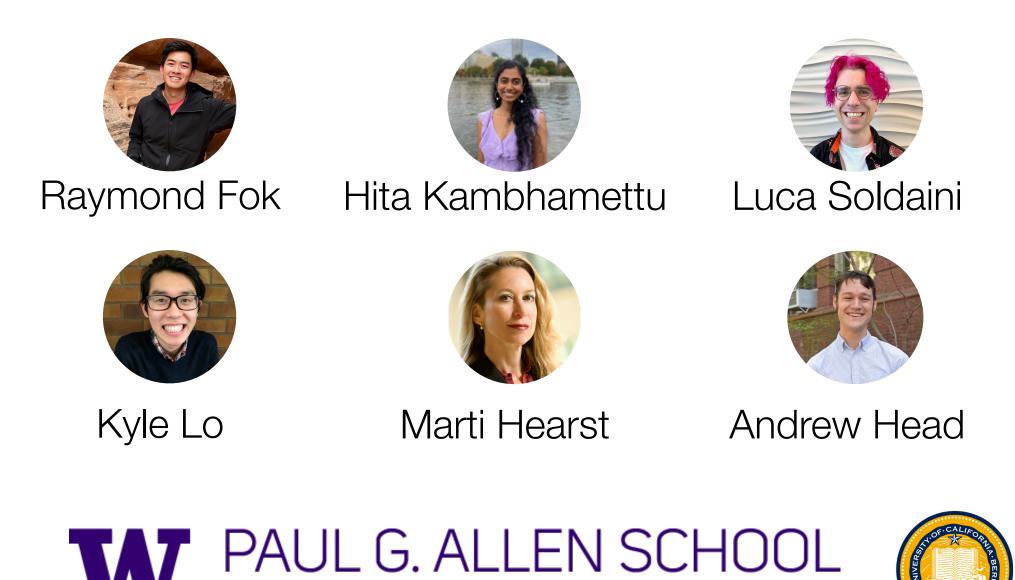
https://www.semanticscholar.org/product/semantic-reader

What's next?

- How should we leverage LLMs (e.g., GPT-4) to support AI-augmented reading?
- How should AI-powered assistants support the creation of academic literature (i.e., writing?)

Scim is an augmented reading interface leveraging intelligent, faceted highlights to support skimming of scientific papers.

Our findings suggest the desire for and promise of AI-powered skimming aids, yet also raise design challenges in augmenting human cognition without introducing undesirable cognitive overload.



OF COMPUTER SCIENCE & ENGINEERING

Summary



Jonathan Bragg



Dan Weld



Paper



Dataset



Code



